

Evaluation of ChatGPT's Implementation in Undergraduate Java Programming Tutorials

Zilin Wang¹, Di Zou^{2*},
Lap-Kei Lee¹, Haoran Xie³, Fu Lee Wang¹
¹ Hong Kong Metropolitan University
² The Hong Kong Polytechnic University
³ Lingnan University
* daisy.zou@polyu.edu.hk

Abstract: This study investigates the impact of integrating ChatGPT into undergraduate Java programming tutorials, focusing on its effects on students' knowledge acquisition, motivation, and self-efficacy. A quasi-experimental design was employed over a seven-week period, involving 53 undergraduate students divided into a ChatGPT-centered learning group (CCLG) and a teacher-centered learning group (TCLG). Both groups completed programming tasks, assessments, and surveys to measure learning outcomes. While the results indicated no statistically significant improvement in programming knowledge, students in the ChatGPT-centered group demonstrated significantly higher motivation and self-efficacy compared to their peers in the traditional instructional setting. These findings suggest that ChatGPT can enhance engagement and confidence in programming education, though its direct impact on knowledge acquisition may be limited within short-term interventions. The study highlights potential challenges, including students' adaptation to AI-assisted learning and limitations in foundational programming skills. Future research should explore strategies for optimizing ChatGPT's implementation, addressing digital literacy gaps, and evaluating its long-term effectiveness in programming education.

Keywords: Programming Education, ChatGPT, Educational Technology, Learning Motivation, Self-Efficacy

1. Introduction

OpenAI's ChatGPT is an advanced language model based on the GPT (Generative Pre-trained Transformer) framework. By utilizing deep learning algorithms, it can interpret and produce human-like text with remarkable accuracy. Trained on an extensive dataset of internet-based content, ChatGPT demonstrates significant potential in educational applications (Liang et al., 2021). Research findings indicate that ChatGPT can enhance student learning outcomes (Cooper, 2023), offer personalized learning pathways (Ghanizadeh et al., 2015), and support collaborative and communicative learning environments (Hinton & Wagemans, 2023). These insights highlight ChatGPT's value as an educational tool.

Despite its potential, assessing ChatGPT's effectiveness within specific educational contexts remains essential. Certain challenges must be addressed, such as concerns over its generalizability and lack of subject-specific expertise risks of misinformation limited human interaction capabilities (Watters & Lemanski, 2023) and ethical issues like its potential misuse for academic dishonesty (Alshurafat et al., 2024). As ChatGPT reshapes traditional educational practices, it becomes increasingly important to explore strategies for optimizing its use in real-world settings.

To contribute to this discussion, this study investigated the impact of ChatGPT in a Java programming tutorial over seven weeks. A total of 53 undergraduate students were divided into two groups: one engaged in ChatGPT-centred learning (CCLG) and the other in teacher-centred instruction (TCLG). Participants completed tutorials, knowledge assessments, and surveys to evaluate their learning achievements, motivation, and self-efficacy.

2. Programming Education

Programming education is essential in higher education, equipping students with computational thinking skills vital for the AI era (Botirovich et al., 2020). However, mastering programming requires complex cognitive abilities, including syntax selection, debugging, problem-solving, and communication (Coşkunserçe, 2023). Traditional teaching methods often emphasize rote learning, leading to low motivation and self-efficacy among students (Botirovich et al., 2020; Konecki, 2014). Teacher-centered learning offers a structured approach by guiding students through coding concepts with direct instruction, enhancing engagement (Elen et al., 2007). However, its effectiveness is limited by the lack of immediate, personalized feedback, particularly during debugging.

To address this, this study explores the use of ChatGPT as a programming assistant. By providing real-time error detection and tailored feedback, ChatGPT can support students in overcoming coding challenges, fostering a more engaging learning experience.

This study investigates:

- (1) Does a ChatGPT-centered approach improve students' knowledge, motivation, and self-efficacy compared to traditional teaching?
- (2) Does prior Java proficiency significantly impact on learning outcomes, motivation, and self-efficacy?

3. ChatGPT-based Java Programming Tutorial

CCLG in this study utilized ChatGPT as a central tool to assist students in completing programming tasks, while the teacher's role was primarily to deliver learning objectives and monitor the class. Figure 1 reports the situation in which the student was using the university ChatGPT web portal to complete the programming task.

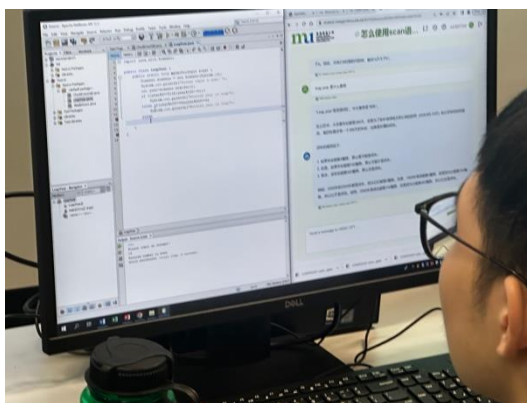


Fig.1 ChatGPT-centered programming learning

Each tutorial followed a structured format consisting of three phases. In the first phase, the teacher introduced the background and context of the programming tasks and explained the objectives in detail. For example, the teacher might assign tasks such as: “Create a game that allows the user to guess an integer number until their answer is correct” or “Create a function that allows the user to determine their grade point when they input a score.” While the teacher introduced the tasks, students quickly listed the programming syntax they would need to use, such as “for loop,” “if statement,” and “switch statement.” This helped students focus on the relevant concepts and tools required to complete the tasks.

In the second phase, students worked on the coding tasks with the support of ChatGPT. They were encouraged to use ChatGPT to translate task instructions into code, define programming syntax they were unfamiliar with, and debug errors in their code. For instance, a student working on the grade-point function task might consult ChatGPT to understand how to implement conditional logic to map score ranges to grade points. During this stage, the teacher moved around the classroom to provide technical support as needed, ensuring that any unresolved issues could be addressed promptly.

The final phase involved a reflective activity designed to deepen students' understanding of both the programming process and their problem-solving strategies. After reviewing the students' completed tasks, the teacher posed reflective

questions, such as: “What problems did you encounter, and how did you overcome them?” “Can you explain your problem-solving logic?” and “Did you use the same solutions as ChatGPT? If not, which solution do you think is better, and why?” These questions encouraged students to critically evaluate their own approaches and compare them to the solutions suggested by ChatGPT, fostering metacognitive skills and a deeper understanding of programming concepts.

4. Methodology

4.1. Participants

Our study initially aimed to recruit 60 sophomore students from a university in Hong Kong. By the end of the experiment, 53 students had completed all the learning activities and tests and were thus included in the final sample. The participants, aged between 18 and 26, were all enrolled in a Java Programming Fundamentals course. According to their responses in the biographical survey, the students had one to two years of programming education experience. However, most had minimal exposure to technology-enhanced learning, AI-based learning, or prompt training.

The participants were randomly assigned to two groups: a control group (TCLG, $N = 27$) and an experimental group (CCLG, $N = 26$).

4.2. Experimental Procedures

In week one, participants completed a 70-minute pre-test assessing their Java knowledge and a 20-minute pre-questionnaire on learning motivation and self-efficacy.

From weeks 2 to 9, students attended weekly two-hour programming tutorials in computer labs. Both groups received identical learning materials but followed different instructional methods. The control group engaged in a teacher-centred approach, where the instructor guided students through tasks, provided explanations, and addressed common coding issues. In contrast, the experimental group utilized a ChatGPT-based learning method.

In week 10, all students took a post-test to assess their learning outcomes and completed post-questionnaires.

4.3. Instruments

This study utilized a pair of programming knowledge tests and questionnaires to evaluate learning outcomes and changes in motivation and self-efficacy. The pre- and post-programming knowledge tests consisted of four questions, with a total maximum score of 50. These included short-answer questions, multiple-choice questions, and code-writing tasks.

The pre- and post-questionnaires were identical and comprised two sections: motivation and self-efficacy. Questions one to six assessed motivation, using items adapted from Wang and Chen (2010) framework on the influence of game strategies on learning motivation. For example, “When I have the opportunity, I choose course assignments that I can learn from even if they don’t guarantee a good grade.” Responses were rated on a five-point Likert scale (1 = strongly disagree, 5 = strongly agree), with higher scores indicating stronger intrinsic and extrinsic motivation.

Questions seven to fourteen measured self-efficacy, defined as an individual’s belief in their ability to perform tasks or achieve goals (Garcia & Pintrich, 2023). This section consisted of eight items adapted from Pintrich’s (1991) self-efficacy questionnaire. An example statement is: “I’m confident I can understand the most complex material presented by the instructor in this course.” Responses were also rated on a five-point Likert scale, with higher scores reflecting greater confidence in programming-related tasks.

We initially conducted reliability analyses and the Cronbach’s Alpha values for both 14-item questionnaires, along with the Cronbach’s Alpha Based on Standardized Items, exceed .8 and approach 90%, indicating that the scales possess high internal consistency. This high level of internal consistency points to a significant degree of reliability in the measurements.

4.4. Data Analysis

At the outset, we conducted a series of Shapiro-Wilk tests to assess the normality of the post-knowledge-test scores and post-questionnaire responses for both groups. The results indicated that only the control group’s post-knowledge-test scores were normally distributed ($p = .377 > .05$). Consequently, we primarily employed

non-parametric tests for data analysis. Firstly, we used Wilcoxon Signed-Rank tests to compare the paired pre- and post-test scores within each group. This allowed us to assess the changes in scores over time for the same participants. Secondly, we employed Mann-Whitney tests to compare the differences between the two groups. Lastly, we utilized Quade's Ranking-Based Covariance Analysis to further confirm the impact of the learning methods on learning performance, while controlling for specific covariates.

5. Results

5.1. Results of Students' Knowledge Test

In this study, pre-test and post-test scores were collected for both groups. The descriptive statistics for these scores are reported in Table 1. The results of the Wilcoxon signed-rank test indicated no significant differences between the pre-test and post-test scores for either group, with both p-values exceeding 0.05. However, it is noteworthy that the mean score of the experimental group demonstrated a slight increase, whereas the control group exhibited a slight decrease.

Table 1. The descriptive result of students' pre- and post-test scores of Java programming knowledge test

	Pre-test		Post-test	
	Mean	SD	Mean	SD
Control group	37.12	6.09	36.38	6.52
Experimental group	31.31	10.08	31.44	10.95

5.2. Results of Students' Motivation

The Wilcoxon signed-rank tests revealed that the experimental group's post-test scores for motivation were significantly higher than their pre-test scores ($p = .000 < .05$), while no significant differences were found in the control group ($p = .762 > .05$). The results of the Mann-Whitney tests further suggested that the experimental group's post-test scores were significantly higher than those of the control group ($Z = -3.311$, $p < .001$). These findings indicate that, compared to the teacher-centered learning method, the ChatGPT-centered learning method significantly improved students' motivation. To further confirm whether the learning method significantly impacted students' motivation scores after controlling for their pre-knowledge-test scores, we conducted Quade's ranking-based covariance analysis. The results, as shown in Table 2, indicated a significant effect of the group variable on post-test motivation scores, $F(1, N-2) = 16.541$, $p < .001$. This suggests that the experimental group demonstrated significantly higher motivation scores compared to the control group after adjusting for pre-test knowledge scores.

Table 2. The Quade's ranking -based covariance analysis for students' post-motivation scores

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	2905.063 ^a	1	2905.063	16.541	< .001***
Intercept	1.034	1	1.034	.006	.939
Learning method	2905.063	1	2905.063	16.541	< .001***
Error	8957.010	51	175.628		
Total	11862.073	53			
Corrected Total	11862.073				

*** $p < .001$

5.3. Results of Students' Self-Efficacy

The Wilcoxon signed-rank tests revealed that the experimental group's post-test scores for motivation were significantly higher than their pre-test scores ($p = .000 < .05$), while no significant differences were found in the control group ($p = .762 > .05$). The results of the Mann-Whitney tests further suggested that the experimental group's post-test scores were significantly higher than those of the control group ($Z = -3.311$, $p < .001$). These findings indicate that, compared to the teacher-centered learning method, the ChatGPT-centered learning method significantly improved

students' motivation. To further confirm whether the learning method significantly impacted students' motivation scores after controlling for their pre-knowledge-test scores, we conducted Quade's ranking-based covariance analysis. The results, as shown in Table 2, indicated a significant effect of the group variable on post-test motivation scores, $F(1, N-2) = 16.541$, $p < .001$. This suggests that the experimental group demonstrated significantly higher motivation scores compared to the control group after adjusting for pre-test knowledge scores.

Table 3. The Quade's ranking -based covariance analysis for students' post-self-efficacy scores

Source	Type III Sum of Squares	df	Mean Square	<i>F</i>	Sig.
Corrected Model	2853.975a	1	2853.975	16.578	< .001***
Intercept	1.016	1	1.016	.006	.939
Learning method	2853.975	1	2853.975	16.578	< .001***
Error	8780.095	51	172.159		
Total	11634.070	53			
Corrected Total	11634.070	52			

*** $p < .001$

6. Discussion and Conclusion

This study investigated the impact of a ChatGPT-centred learning method on students' knowledge acquisition, motivation, and self-efficacy in Java programming education. While the results revealed a significant improvement in motivation and self-efficacy for students in the experimental group, the impact on knowledge acquisition was minimal.

The insignificant differences in students' pre-test and post-test scores suggest that while ChatGPT enhanced engagement and confidence, it did not substantially improve their coding proficiency within the study period. One possible explanation is the small sample size (Carlson & Schmidt, 1999), which may have limited the ability to detect statistically significant improvements. In studies with small groups, minor variations in performance can be amplified, leading to non-significant findings even when an actual effect exists. Future studies with larger sample sizes could help clarify whether ChatGPT meaningfully enhances programming knowledge acquisition.

Another contributing factor may be students' limited foundational digital skills. A lack of prior digital literacy can hinder learners' ability to navigate programming environments effectively, thereby reducing their ability to leverage ChatGPT optimally (Mokaya et al., 2022). Students with lower self-efficacy in technology-related domains may struggle with programming concepts, leading to minimal measurable improvement in knowledge despite increased motivation. Addressing these gaps through introductory digital literacy training before engaging with AI-based tools may enhance the effectiveness of such interventions.

Additionally, students' familiarity with traditional teacher-centred learning approaches could have influenced the results. Many learners may have been accustomed to direct instruction and guided problem-solving rather than self-directed learning facilitated by ChatGPT. This shift in instructional methods may have required an adaptation period, limiting immediate knowledge gains. Some students may also have held negative attitudes toward ChatGPT, either due to skepticism about AI-generated feedback or discomfort in relying on a non-human instructor. Such resistance could have affected engagement with the tool and subsequently impacted their learning outcomes. Future studies should explore strategies to gradually integrate ChatGPT into programming education, ensuring sufficient guidance and scaffolding to support students in transitioning to AI-assisted learning.

Despite these limitations, the study aligns with prior research indicating that generative AI enhances motivation and self-efficacy (Mun, 2024; Woo et al., 2024). ChatGPT's ability to provide instant, contextually relevant feedback creates a more interactive learning experience, reducing anxiety and fostering a sense of achievement (Mun, 2024). By minimizing students' fear of failure, ChatGPT enables them to engage with programming in a more enjoyable and stress-free manner (Wu et al., 2024). In terms of self-efficacy, students in the experimental group exhibited greater

confidence in their programming abilities, possibly due to the ease of accessing ChatGPT's explanations and debugging assistance. When students can quickly identify errors and receive immediate feedback, they develop a stronger belief in their ability to solve problems independently (Mun, 2024). However, the challenge remains in ensuring that this increased confidence translates into measurable improvements in programming proficiency.

This study also has limitations. The difficulty level of the knowledge tests may have contributed to low scores, making it harder to capture improvements in students' learning outcomes. Future research should consider adjusting the complexity of assessments to better align with students' abilities. Additionally, the absence of qualitative data limits deeper insights into students' perceptions of ChatGPT-assisted learning. Incorporating interviews or focus groups could provide a more comprehensive understanding of how learners interact with AI tools and what challenges they face.

Future research should address these gaps by incorporating larger sample sizes, refining assessment methods, and integrating qualitative research to capture student experiences more holistically. Additionally, exploring long-term effects and cross-disciplinary applications of ChatGPT in education could provide further evidence of its potential as a learning tool. By refining AI-assisted instructional strategies, educators can better leverage ChatGPT to enhance programming education and support diverse learners effectively.

Acknowledgements

We thank all the people who wrote previous versions of this document.

References

- Alshurafat, H., Al Shbail, M. O., Hamdan, A., Al-Dmour, A., & Ensour, W. (2024). Factors affecting accounting students' misuse of chatgpt: an application of the fraud triangle theory. *Journal of Financial Reporting and Accounting*, 22(2), 274–288.
- Botirovich, B. D., Ergashevich, T. K., Eshmirzayevna, M. U., Kholboyevich, A. S., Fayzullayevna, K. N., & Mukhiddinovna, A. U. (2020). The importance of teaching algorithms and programming languages in the creation of electronic education resources. *Journal of Critical Reviews*, 7(11), 365–368.
- Carlson, K. D., & Schmidt, F. L. (1999). Impact of experimental design on effect size: Findings from the research literature on training. *Journal of Applied Psychology*, 84(6), 851.
- Cooper, G. (2023). Examining science education in chatgpt: An exploratory study of generative artificial intelligence. *Journal of Science Education and Technology*, 32(3), 444–452.
- Coşkunserçe, O. (2023). Comparing the use of block-based and robot programming in introductory programming education: Effects on perceptions of programming self-efficacy. *Computer Applications in Engineering Education*, 31(5), 1234–1255.
- Elen, J., Clarebout, G., Léonard, R., & Lowyck, J. (2007). Student-centred and teacher-centred learning environments: What students think. *Teaching in Higher Education*, 12(1), 105–117.
- Garcia, T., & Pintrich, P. R. (2023). Regulating motivation and cognition in the classroom: The role of self-schemas and self-regulatory strategies. In *Self-regulation of learning and performance* (pp. 127–153). Routledge.
- Ghanizadeh, A., Razavi, A., & Jahedizadeh, S. (2015). Technology-enhanced language learning (TELL): A review of resources and upshots. *International Letters of Chemistry, Physics and Astronomy*, 54, 73–87.
- Hinton, M., & Wagemans, J. H. M. (2023). How persuasive is AI-generated argumentation? An analysis of the quality of an argumentative text produced by the GPT-3 AI text generator. *Argument & Computation*, Preprint, 1–16. <https://doi.org/10.3233/AAC-210026>
- Konecki, M. (2014). Problems in programming education and means of their improvement. *DAAAM International Scientific Book*, 2014, 459–470.

- Liang, J.-C., Hwang, G.-J., Chen, M.-R. A., & Darmawansah, D. (2021). Roles and research foci of artificial intelligence in language education: an integrated bibliographic analysis and systematic review approach. *Interactive Learning Environments*, 1–27.
- Mokaya, M., Kyallo, F., Vangoitsenhoven, R., & Matthys, C. (2022). Clinical and patient-centered implementation outcomes of mHealth interventions for type 2 diabetes in low-and-middle income countries: a systematic review. *International Journal of Behavioral Nutrition and Physical Activity*, 19(1), 1.
- Mun, I. B. (2024). A study of the impact of ChatGPT self-efficacy on the information seeking behaviors in ChatGPT: the mediating roles of ChatGPT characteristics and utility. *Online Information Review*.
- Pintrich, P. R. (1991). A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ).
- Wang, L., & Chen, M. (2010). The effects of game strategy and preference-matching on flow experience and programming performance in game-based learning. *Innovations in Education and Teaching International*, 47(1), 39–52.
- Watters, C., & Lemanski, M. K. (2023). Universal skepticism of ChatGPT: a review of early literature on chat generative pre-trained transformer. *Frontiers in Big Data*, 6, 1224976.
- Woo, D. J., Wang, D., Guo, K., & Susanto, H. (2024). Teaching EFL students to write with ChatGPT: Students' motivation to learn, cognitive load, and satisfaction with the learning process. *Education and Information Technologies*, 1–28.
- Wu, T.-T., Lee, H.-Y., Li, P.-H., Huang, C.-N., & Huang, Y.-M. (2024). Promoting self-regulation progress and knowledge construction in blended learning via ChatGPT-based learning aid. *Journal of Educational Computing Research*, 61(8), 3–31.